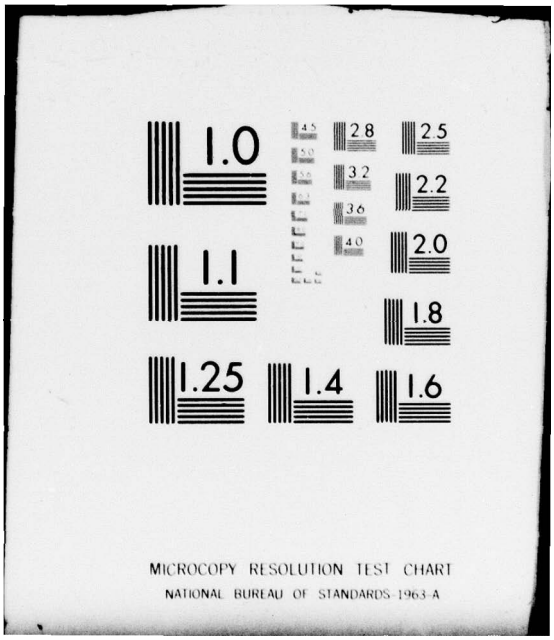
AD-A074 842 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
PEER AND SUPERVISORY RATINGS OF RESEARCH SCIENTISTS.(U)
SEP 79 G D KISSLER, D M NEBEKER

UNCLASSIFIED NPRDC-TR-79-31 NL

| OF |
AD A074842

END
DATE FILMED
11-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

NPRDC TR 79-31

September 1979

PEER AND SUPERVISORY RATINGS OF RESEARCH SCIENTISTS

Gary D. Kissler
Delbert M. Nebeker

Reviewed by
Richard C. Sorenson

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: NPRDC-TR-79-31	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): PEER AND SUPERVISORY RATINGS OF RESEARCH SCIENTISTS,	5. TYPE OF REPORT & PERIOD COVERED: Final Report, for FY79
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Gary D. Kissler, Delbert M. Nebeker	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Navy Personnel Research and Development Center, San Diego, California 92152	12. REPORT DATE: Sep[illegible] 1979 / 13. NUMBER OF PAGES: 24
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED / 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Peer rating
Performance evaluation
Scientific productivity
Supervisory rating

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A sample of 103 government research scientists was used to compare two performance evaluation systems currently used in a Federal agency for its research personnel. The two systems, supervisory ratings and peer ratings, were compared in terms of their respective reliability and validity. The results showed the peer ratings to be more stable over time and to relate more highly to scientific "productivity" than do the supervisory ratings. Also, productivity was found to be significantly related to

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

occupational levels resulting from peer evaluations. A discussion of these results and possible explanations for the differences between the two evaluation processes are given along with other considerations for organizations that contemplate alternate evaluation processes similar to peer ratings.

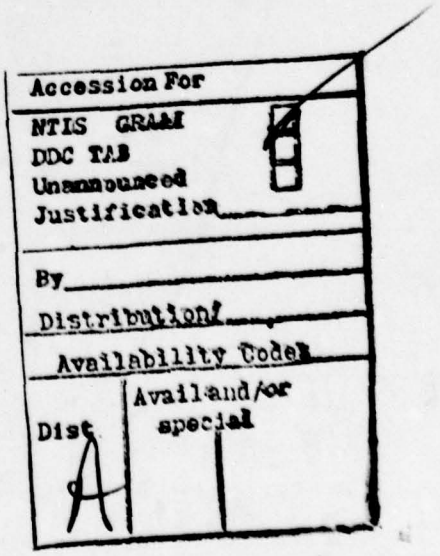
FOREWORD

Reliable and valid methods of employee evaluation are required if organizational rewards are to be used successfully in improving the motivation and performance of Navy personnel. This study evaluated two alternative techniques for personnel performance evaluation.

The results of the work reported here are intended primarily for use by Navy and Federal organizations that employ significant numbers of research scientists, in particular, the Navy Laboratories and Centers.

Appreciation is expressed to the many individuals in the Agricultural Research Service, Department of Agriculture who contributed to and participated in this effort. The data for this project were gathered by Dr. Gary D. Kissler while he was employed by the Department of Agriculture.

DONALD F. PARKER
Commanding Officer



SUMMARY

Problem

Individual evaluation in organizations has developed along two major lines, rating scales and alternate assessment sources. A long history of disappointing efforts to develop efficient and useful rating scales exists. Among alternate evaluative methods, no in-place peer-evaluation processes that affect personnel actions have received attention in the literature. A recent comprehensive review of peer-assessment (Kane & Lawler, 1978) concluded that none of the studies reported had included an adequately objective measure of performance and that this issue needs to be addressed through the use of multimethod-multisource designs.

Objective

This report compares the psychometric properties of peer and supervisory ratings being used to evaluate Government research scientists within a Federal agency. These two measurement systems were compared by their stability over time and their validity. For the latter comparison, each set of ratings was related to measures of scientific performance, e.g., honors and awards received and reports and articles published. Interrater reliability for peer ratings was also calculated. The basic hypotheses were: (1) peer ratings are more reliable than supervisory ratings and (2) peer ratings are more highly related to scientists' "productivity" than are supervisory ratings.

Approach

A group of Government agricultural scientists (n = 103) was chosen for this study. During a 6-month period, each of the scientists in this study was scheduled for a position review by a peer panel. Objective measures of scientific performance included number of honors and awards received, number of science-oriented meetings attended, frequency of advising or consulting activities, number of special invitations to present research findings, number of publications, and frequency of individual research citations over a 3-year period. Current and previous peer ratings were taken from a process used to assess the positions of over 4000 scientists. Supervisory ratings were also obtained from agency records for 3 consecutive years.

Findings

1. Current and previous peer ratings are highly similar, suggesting that peer ratings remain stable over time.

2. The peer panels show high rating agreement among the members prior to discussions.

3. Current and previous supervisory ratings are not as similar over time as are the peer ratings.

4. Peer ratings are highly related to all performance measures, especially number of publications. They show a relatively low relationship to the frequency of advising activity.

5. Supervisory ratings are not as highly related to performance measures as are peer ratings. Like the peer ratings, however, they are most strongly related to number of publications and least strongly related to frequency of advising activities.

Conclusions

The peer rating system shows more stability over time than the supervisory rating system. There is substantial agreement among peer raters concerning the evaluation of scientific "productivity."

The peer ratings show stronger evidence of validity than do the supervisory ratings.

Recommendations

1. An organization faced with the task of evaluating performance, particularly performance by higher level employees such as those participating in this study, should consider the merits of implementing a system whereby peer input could be used.

2. The consistency, accuracy, and fairness demonstrated by the peer rating system used in this study suggests that organizations employing such a process will be better able to justify personnel decisions and could encourage high productivity by making the connection between individual performance and organizational rewards clearer through more objective performance criteria.

CONTENTS

INTRODUCTION

Problem and Background

Efforts to improve the evaluation of individuals within organizations have developed along two major lines. The first seeks to improve validity by improving rating methods, while the second emphasizes alternate assessment sources.

A long history of efforts to develop efficient and useful rating scales exists (Campbell, Dunnette, Lawler, & Weick, 1970). One effort to improve ratings has focused on altering rating formats. For example, some researchers (Barrett, Taylor, Parker, & Martens, 1958; Smith & Kendall, 1963) have proposed behaviorally anchored rating scales that substitute behavioral descriptions (e.g., satisfactory or unsatisfactory) for evaluative descriptions. Despite early claims supporting this approach, recent research (Borman & Dunnette, 1975; Bernardin, Alvares, & Cranny, 1976; Bernardin, 1977; DeCotiis, 1977) has not demonstrated the superiority of these scales over more traditional measures. (For a more complete review, see Schwab, Heneman, & DeCotiis, 1975.) Another example of format change is found in the job satisfaction area when verbal scale descriptions are replaced with a series of male faces ranging from a broad smile to a deep frown to measure job satisfaction (Kunin, 1955). Although evidence exists to support the use of such a scale (Locke, Smith, Kendall, Hulin, & Miller, 1964; Smith, Kendall, & Hulin, 1969), including one with female faces (Dunham & Herman, 1975), ironically, such scales may lack "face validity" when used with different occupational levels.

It would appear that modifying rating scales does not yield much improvement. Burnaska and Hollman (1974) draw a similar conclusion:

> . . . we see . . . that format modification and improvement in scale development techniques are not the final remedy to the criterion problem. Other approaches to improve ratings . . . may prove to be equally or more fruitful alternatives. (p. 312)

The second line of research seeks out the "best" of several information sources instead of changing rating scales used by a predetermined source. Smith (1976) states that alternate performance evaluation sources include self, subordinate, superior, and peer ratings. The present study compares the last two sources.

The earliest reported use of peer ratings involved military personnel (Williams & Leavitt, 1974; McClure, Tupes, & Dailey, 1951) and occurred several years before peer ratings were considered for an industrial setting (Weitz, 1958; Roadman, 1964). Reviews of literature (Korman, 1968; Lewin & Zwany, 1976) have documented fairly impressive progress for this approach in terms of stability, predictive validity, and resistance to bias. Lawler (1967) found evidence of good convergent validity and moderate discriminant validity between peer and superior ratings of managers in a manufacturing organization. He concluded that the extent to which rating scales require raters to make judgments rather than evaluate actual performance may influence the degree of improvement expected from alternative sources. Peer ratings based on closer proximity to the performance being examined have the potential to improve evaluation. It is necessary to stress the word "potential," since the previous civilian research (Weitz, 1958; Roadman, 1964; Waters & Waters, 1970; Mayfield, 1972) has involved only tests of peer ratings as possible evaluation systems. No in-place peer-evaluation processes that affect personnel actions have received attention in the literature. As Smith (1976) also points out, few civilian applications exist in this area. Lawler (1967) states that

> . . . perhaps the most important problem with respect to the use of peer ratings for personnel decision making is the problem of the research set versus the administrative set . . . if the rater knows it is "going to count," ratings may lose their validity . . . . (p. 379)

More recently, a comprehensive review of peer assessment methods by Kane and Lawler (1978) examined the psychometric and administrative properties of peer ratings. None of their cited studies, however, attempted to compare these properties directly for supervisory and peer ratings. Further, their review states that

> . . . none of the studies reported to date has included an adequately objective measure of performance . . . (and a) clear need exists to address this issue through the use of multimethod-multisource designs. (p. 512)

Purpose

This report compares the psychometric properties of two systems currently used to evaluate research scientists within a Federal agency. The first is a fairly typical supervisory rating system; and the second, a peer rating system. The peer evaluation process differs by design from some peer rating methods to enhance its usefulness in an applied setting. The comparison focuses on the systems of evaluation, since it is acknowledged that the source and associated method of ratings are confounded to some degree. Specifically, the rating methods are examined to compare their stability over time and their validity. To determine validity, each set of ratings will be related to several different measures of scientific performance, for example, honors and awards, and publications. For peer ratings, conspect reliability will also be calculated. Our basic hypotheses are (1) that the peer ratings are more reliable than supervisor ratings and (2) that the peer ratings are more highly related to scientists' "productivity" than supervisory ratings.

METHOD

Subjects

A group of Government agricultural scientists (n = 103) was chosen for this study. These researchers were mostly male Caucasians who ranged in grade from GS-11[1] to GS-16 ($\bar{X}$ = 13.3; SD = 1.5), in age from 29 to 66 years ($\bar{X}$ = 48; SD = 8.8), in length of service from less than 1 year to 40 years ($\bar{X}$ = 15.9; SD = 8.5), and in time in present grade from less than 1 year to 16 years ($\bar{X}$ = 7.2; SD = 3.3). Seventy-eight percent had attained the doctoral degree. Agency records were combined with information pertaining to shared research interests, projects, and activities to classify these individuals into six research peer groups: animal (n = 19) and plant (n = 28); food, chemistry, and nutrition (n = 20); entomology (n = 18); environmental processing and mechanical engineering (n = 9); and soil science (n = 7). Two remaining scientists belong to the food marketing peer group, which is unique and therefore not combined with other groups. Results requiring the designation of separate peer groups will be discussed later.

During a 6-month period, each of the scientists in this study was scheduled for a position review by a peer panel. Ten different panels were formed and each rated approximately ten scientists. Since random selection was not possible, the extent to which the numbers of scientists in each peer group proportionately represent the total numbers in the agency peer groups was examined. A Chi-square test of the proportional differences found nonsignificant results (X^2 = 6.86; df = 5) and, furthermore, no selective factors affecting order of evaluation could be found.

Measures

Objective Performance Measures

A mandatory peer review takes place for the scientists in this organization every 3 to 5 years from the date of hire, with more frequent reviews available at the option of the scientist. Each researcher prepares the materials that reviewers will use to evaluate his or her performance. The format and content of these materials are outlined in advance to minimize style differences. Six areas of information are requested: (1) research assignment, (2) extent of supervision received, (3) examples of supervisory and team leadership, (4) documented research accomplishments, (5) work-related biographical data, and (6) unusual accomplishments or circumstances involving the scientist. In addition, each scientist provides a list of his publications, noting those published since the date of the last promotion. These measures are not abstracted for the reviewers, since most of the information is in narrative form. By examining each set of materials provided for the panel, however, it was possible to quantify the following measures of performance.

1. Number of honors and awards received.
2. Number of science-oriented meetings attended.
3. Frequency of advising or consulting activities.
4. Number of special invitations to present research findings.
5. Number of publications.

---

[1]Information on U.S. Government grades and pay plans can be found in a Federal Personnel Manual available from the U.S. Civil Service Commission, Washington, DC.

In addition, a sixth objective measure of the scientist's performance was obtained by evaluating the impact of each scientist's research on the scientific community (Lawani, 1977). The Science Citation Index was used to tabulate the frequency of individual research citations over a 3-year period. This measure was not part of the personal review process but was monitored as an additional check on the validity of the performance measurement systems. A log transform of each performance variable was taken to normalize a tendency toward positively skewed distributions.

Peer Ratings

In the late 1950's a collective effort on the part of research scientists and personnel specialists employed by the agency from which the present sample was drawn resulted in a proposed peer evaluation process that is currently used to assess the positions of over 4000 agricultural scientists (Note 1; Note 2). This process was reviewed by the Civil Service Commission and received some preliminary evaluation (McKean, Mandel, & Steel, 1960) before being slightly revised and authorized for use throughout the Federal Government (Note 3). The process will be discussed in three sections: panel membership, evaluation process, and use of output.

Each peer evaluation panel consists of seven members, six scientists and a personnel classification specialist. A minimum of two scientists must represent the peer group of each researcher being evaluated. One of these two scientists acts as an indepth reviewer charged with the responsibility of seeking out any additional information not included in the written materials that would aid the evaluation process. The panel is chaired by one of the six scientists. Prior to meeting, each panel member receives and independently evaluates the written materials on four factors:

1. Type of research assignment. The responsibilities assigned to the researcher, his research objectives and methodology, and the results expected of him are considered.

2. Supervision received. The authority assigned to the researcher, the type of research guidance he receives, the degree to which his results are reviewed, and the amount of general supervision he receives are considered.

3. Guidelines and originality. The amount of literature available to the scientist in the major area of research, the extent to which originality is required in this area, and the degree to which he demonstrates originality are considered.

4. Qualifications and contributions. Each scientist's demonstrated research stature, scientific recognition, impact on science and technology, and advisory or consulting activities are considered.

A fairly high degree of correspondence is expected between ratings of these factors, since it is recommended that supervisors match the research task and degree of supervision to be received by their subordinates to their qualifications and contributions as well as their originality. The process of evaluation begins as the panel members independently examine each scientist's prepared materials. Each member is provided standard agency forms that contain five written descriptions of scientific activities corresponding to levels of performance for each of the criteria described above. Each level is associated with a number ranging from 2 to 14 for the first three criteria; the fourth receives double weighting to offset undue emphasis upon assignment and work situations. The panel member assigns a numerical score for each criterion on the basis of the perceived similarity between the written materials and written descriptions of performance. The sum of these scores is also recorded for each individual. Although the

preparation and distribution of materials is structured, panel members are not given specific rules for converting the materials to numerical scores, nor are they told how the various elements of the materials should be weighted in determining ratings. Once each panel member has independently evaluated each scientist, the panel is ready to meet.

The overall objective of the panel meeting is to achieve a consensus across the four factors for each researcher's position. To begin, each panel member is asked to present his or her scores for the scientists. Following this, and prior to discussion, the indepth reviewer presents any additional information, some of which could be obtained from the supervisor. By requesting and exchanging information, the members eventually agree. In rare instances, the panel cannot agree and an arbitrator from upper management is asked for an opinion. This did not occur in the cases reported here.

Based on established numerical point-distributions, the decision of the panel is translated into one of three personnel actions: (1) retention in grade, (2) promotion to next higher grade, or (3) demotion to next lower grade. The final responsibility of the panel is to explain to the scientist the basis for its decision on each criterion. For each researcher in this study, it was possible to obtain the panel scores for each criterion. It is interesting to note that the panels are newly composed at each review so that the scores from any previous panel represent the evaluations of a different group of peers. For most of the scientists, scores from the last previous panel review were also available to us but not to the review panel.

Supervisory Ratings

Government regulations require that Federal employees receive yearly performance ratings from their supervisors. For the present sample, this was accomplished by completing an agency form that contained numbered evaluative descriptions, ranging from 1 (unsatisfactory) to 9 (distinguished), that are applied to three categories: quantity of work, quality of work, and supervising others. In this evaluation process, the supervisors have access to the same material used by the peer panels.

Agency records provided supervisory ratings for 3 consecutive years for most scientists. Supervisory ratings generally affect personnel actions other than promotion, for example, awards and reduction in force actions. Any input from the supervisor in the peer rating process comes verbally through the indepth reviewer and at no time are the supervisor's numerical ratings considered by the peer panel. Neither peer ratings nor supervisory ratings incorporate information regarding research citations, the sixth objective measure of performance.

RESULTS

Reliability Measures

Reliability for the peer ratings was computed by test-retest and internal agreement methods. Stability coefficients for each peer-rating criterion measured during the current review and the previous review are located in the upper left-hand triangle of Table 1. The current and previous criterion coefficients range from .89 to .92. The current and previous total scores across the criteria correlated .95. The range for all peer rating coefficients is .86 to .96 and all are significant at the $p < .001$ level. It is important to remember that these reliabilities are based on evaluations from different peer review panels 3 to 5 years apart. Lest one assume that the panels were merely duplicating the evaluations of the previous panels, it should be noted that the previous panel's ratings are not available to the current panel members.

Interrater agreement was assessed by using an analysis of variance (Winer, 1971; p. 283) among peers across the four criteria, making it possible to examine how closely the panel members agreed prior to discussion. For each scientist, an intraclass coefficient was calculated. The mean, standard deviation, and median of these figures are .96, .06, and .99, respectively, indicating exceptionally high similarities among the independent evaluations. Both of these methods of estimating reliability suggest that the peer rating system is highly reliable.

Table 1 also contains the stability coefficients for the supervisory ratings in the lower right-hand triangle. To provide figures comparable to those for peer ratings the coefficients are displayed representing two supervisory ratings given approximately 2 years apart. The current and previous coefficients for the categories "quantity," "quality," and "supervising others" are .63, .69, and .58, respectively. An average was calculated across the ratings each scientist received in these categories for each of the two time periods. These averaged ratings correlated .71. The range for all coefficients in the triangle is .41 to .76, all significant at the $p < .001$ level. While it appears that the reliability of the supervisory ratings is often substantial for the 2-year period, it is considerably less than that found for the peer evaluation.

Validity Measures

Three points should be made prior to presenting results pertaining to validity. First, the lower rectangle of Table 1 shows the relationships between peer and supervisory ratings. These coefficients show a fair degree of convergent validity between the rating systems, and they are somewhat higher than those reported by Kane and Lawler (1978). This superiority is probably due, in part, to the higher reliabilities found in the present study.

Second, as can be seen in Table 1, the elements of each rating system demonstrate substantial overlap. While the following results will include separate treatment of these elements, the authors feel the total score and previous total score for peer ratings as well as the averaged supervisory ratings represent accurate composite variables for each type of evaluation. Therefore, both the individual and composite variables will appear in tabular form.

Third, following Dunnette's (1963) suggestion that performance is most accurately considered as multidimensional, validity results are presented that examine several measures of scientific performance. An additional effort was made to create a composite

Table 1

Intercorrelations Among Peer and Supervisory Ratings

	Factor 1	Factor 2	Factor 3	Factor 4	Total Score	Prev. Factor 1	Prev. Factor 2	Prev. Factor 3	Prev. Factor 4	Prev. Total Score	Quantity	Quality	Supv.	Average	Prev. Quantity	Prev. Quality	Prev. Supv.	Prev. Average	M	SD
Peer Ratings[a]																				
Factor 1	1.0																		7.2	2.7
Factor 2	.92	1.0																	7.5	2.8
Factor 3	.93	.93	1.0																6.9	2.8
Factor 4	.92	.93	.93	1.0															13.7	5.9
Total Score	.96	.97	.97	.98	1.0														35.3	13.8
Prev. Factor 1	.90	.88	.90	.88	.92	1.0													7.1	2.7
Prev. Factor 2	.91	.89	.91	.88	.92	.92	1.0												7.2	2.7
Prev. Factor 3	.90	.88	.90	.89	.92	.91	.92	1.0											7.0	2.8
Prev. Factor 4	.86	.87	.89	.92	.92	.89	.89	.90	1.0										12.8	6.5
Prev. Total Score	.92	.91	.93	.93	.95	.95	.96	.96	.98	1.0									34.2	14.2
Supervisory Ratings																				
Quantity	.52	.44	.48	.50	.50	.36	.33	.40	.41	.40	1.0								6.7	1.0
Quality	.44	.44	.40	.48	.46	.35	.33	.37	.43	.40	.75	1.0							6.8	1.0
Supv.	.40	.35**	.32**	.42	.40	.35*	.25*	.36**	.34**	.34**	.56	.64	1.0						6.4	1.2
Average	.44	.41	.36	.43	.42	.30**	.28**	.34	.37	.35	.77	.85	.81	1.0					6.6	1.0
Prev. Quantity	.42	.39	.39	.50	.46	.41	.35	.35	.44	.42	.63	.62	.53	.57	1.0				6.5	1.2
Prev. Quality	.53	.54	.53	.58	.57	.55	.51	.52	.58	.57	.75	.69	.41	.65	.76	1.0			6.7	1.0
Prev. Supv.	.47	.46	.46	.55	.52	.56	.48	.49	.53	.54	.53	.47	.58	.60	.76	.68	1.0		6.4	1.1
Prev. Average	.47	.48	.45	.54	.51	.50	.43	.44	.54	.51	.65	.66	.57	.71	.88	.83	.85	1.0	6.6	1.0

Note. (1) n ranged 66-89 for validity coefficients; (2) n ranged 89-103 for peer rating intercorrelations; (3) n ranged 71-89 for supervisory rating intercorrelations.

All coefficients are significant at $p < .001$ unless otherwise noted.

[a]The time between current and previous peer ratings varied between 3 and 5 years, while that between current and previous supervisory ratings was approximately 2 years for all scientists studied.

*$p < .05$.
**$p < .01$.

measure of performance, however. This measure was based on weights generated by analyzing the objective performance variables. A principal components analysis resulted in a single factor accounting for 50 percent of the total variance. The resulting variable, hereafter called "total," is included with the other performance measures.

The top half of Table 2 contains the validity coefficients for the peer rating factors as well as the composite totals. It can be seen that all of these relationships are significant at the $p < .001$ level. Of all the individual performance measures, the number of publications tends to show the highest relationship to peer ratings. The range of coefficients is .59 to .70. The frequency of advising activity shows the lowest overall relationship to peer ratings; the range of coefficients for this variable is .36 to .50. The composite variable "total" shows substantial relationship to peer ratings and has coefficients ranging from .48 to 64. It is important to note that performance as measured by citations was consistently related to peer evaluations even though it was not used in the review process.

Validity coefficients for supervisory ratings are found in the bottom half of Table 2. As with peer ratings, supervisory ratings are most highly related to number of publications. The coefficients range from .20 to .47. The performance variable showing the lowest relationships to supervisory ratings is the frequency of advising activity; the coefficients range from .10 to .23. The variable "total" showed statistically significant relationships to these ratings in most instances; the values range from .10 to .32. It is clear that these validity coefficients are not only smaller than those for peer ratings but many fall below conventional levels of statistical significance. Even when these validity coefficients are corrected for attenuation due to unreliability, they are still inferior to those for the peer evaluations.

A final point regarding validity can be made by returning to Table 1. Here one can reexamine the interrelationships mentioned earlier with the knowledge of which rating system is more closely associated with research "productivity." Given that each system shows relatively high intrarating relationships, that the two systems have moderate but significant intercorrelations, and that the peer ratings relate more highly than the supervisory ratings to measures of scientific performance, one can draw the inference that peer ratings are in general agreement with supervisory ratings but are superior in terms of their relationships to the work done by researchers. Therefore, the intercorrelations of two independent sets of judgments provide additional evidence of validity for each rating system, and correlations between performance measures and each of these systems offer stronger evidence of validity for the peer ratings.

Two criticisms that may be made of the peer rating approach to performance evaluation are (1) that such a system may be compromised by one or more peer groups being more lenient for their own members and, (2) that peer ratings given to the individual may not be directly related to actual productivity but are supported by the current grade level of the scientist regardless of the individual's record of research activity. The first issue was examined by cross-tabulating peer-group by personnel action (promote versus remain in grade). A Chi-square test of the independence of these categories indicated that peer groups were not treated differently by the different peer panels ($X^2 = 3.77$; $df = 5$; ns). To respond to the second criticism, the grade levels of the scientists were correlated with each individual's performance across the objective measures. Grade correlated significantly with honors and awards ($r = .24$; $p < .01$), number of scientific meetings ($r = .51$; $p < .001$), advising and consulting ($r = 46$; $p < .001$), special invitations ($r = .66$; $p < .001$), number of publications ($r = .66$; $p < .001$), number of citations ($r = .51$; $p < .001$), and "total" ($r = .52$; $p < .001$). In each case, a positive and significant relationship exists between research "productivity" and grade level.

Table 2

Correlations Between Peer and Supervisory Ratings and Performance Measures

Ratings[a]	Performance Measures: Honors	Meetings	Advising	Special Invitations	Publications	Citations[a]	Total[b]
Peer Ratings							
Factor 1	.58***	.45***	.41***	.65***	.65***	.49***	.57***
Factor 2	.58***	.50***	.44***	.66***	.68***	.50***	.58***
Factor 3	.59***	.44***	.42***	.69***	.65***	.52***	.60***
Factor 4	.59***	.49***	.50***	.68***	.70***	.51***	.64***
Total	.60***	.49***	.47***	.69***	.70***	.52***	.62***
Previous Factor 1	.55***	.44***	.37***	.61***	.63***	.44***	.49***
Previous Factor 2	.60***	.43***	.37***	.64***	.64***	.46***	.53***
Previous Factor 3	.57***	.44***	.36***	.61***	.59***	.48***	.48***
Previous Factor 4	.58***	.40***	.49***	.63***	.67***	.48***	.55***
Previous Total	.60***	.44***	.43***	.65***	.67***	.48***	.54***
Supervisory Ratings							
Quantity	.35***	.26**	.15	.26**	.37***	.23*	.32***
Quality	.31***	.31**	.23*	.31***	.35***	.30**	.35***
Supv.	.20*	.17	.10	.20*	.21*	.20*	.10
Average	.23*	.27**	.13	.25**	.30**	.24*	.19*
Prev. Quantity	.25**	.15	.15	.19*	.39***	.23*	.20*
Prev. Quality	.35***	.22	.22*	.31**	.47***	.42***	.26**
Prev. Supv.	.25*	.19	.14	.24*	.44***	.27**	.21*
Prev. Average	.27**	.19*	.21*	.26**	.45***	.27**	.25*

Note. (1) n ranged 88-103 for coefficients involving peer ratings; (2) n ranged 70-89 for coefficients involving supervisory ratings.

[a]The time between current and previous peer ratings varied between 3 and 5 years, while that between current and previous supervisory ratings was approximately 2 years for all scientists studied.

[b]This information was not available to peer or supervisor raters.

[c]This variable as used was not available to raters but, instead, is a linear composite of the various sources of information.

*$p < .05$.

**$p < .01$.

***$p < .001$.

DISCUSSION

The most critical hypothesis tested in this paper was that peer ratings are more highly related to objective measures of scientific productivity than are supervisory ratings. Clearly, this was shown to be so whether one was examining each of several measures of research performance or a composite measure of these variables. These findings are of even more interest when one considers that they are based on performance evaluations that show high degrees of reliability.

Test-retest reliabilities for both rating systems indicate that the scientists receive consistent performance evaluations over time, although the amount of overlap between peer ratings appeared greater than that for supervisory ratings. Independent raters in the peer rating system demonstrated remarkable agreement across evaluation criteria prior to group discussions. These interrater reliabilities indicate that the value of scientific performance is perceived in similar ways by individuals separated by distance and perhaps also by discipline or occupation. Validity coefficients based on relationships between each rating system and several measures of research performance demonstrate that each system shows generally statistically significant correlations with performance. Peer ratings, however, are all significantly related to performance while supervisory ratings show lower and, in several instances, nonsignificant coefficients. When the two systems are interrelated, they show moderate and significant relationships, indicating convergent validity among the ratings, despite the superiority of peer ratings. Lawler's suggestion (1967) that peer ratings may lose their validity when they are used for personnel actions does not receive support from our findings. On the contrary, they appear to be quite valid and to warrant the reliance placed upon them by the organization this study investigated.

There could be many reasons why the peer ratings were found to be superior to supervisory ratings. While this research was not designed to explain the differences between the two systems, some discussion of them seems warranted. First, the peer evaluation process in this instance is highly structured. One must admit that few supervisory ratings are based on instructions as specific as those given to the peer raters. Further, the final evaluation receives input from several individuals rather than only one. Even though the interrater agreement was exceptionally high prior to discussion, the anticipation of this discussion may have resulted in a more thorough performance evaluation. The degree of structure in this system must be considered in generalizing these results.

Some may argue that the procedures followed by the peer review panels and supervisors were sufficiently different to make comparison difficult. It should be pointed out, however, that the purpose of both systems was the same--the evaluation of the performance of research scientists. Moreover, while the peer review process was more structured than the supervisory rating system, both were considered important activities and were taken seriously by the individuals involved. These two systems are interesting because they are both operating simultaneously in the same organization. It is unusual to find two personnel performance evaluation systems being used routinely, especially when one of the systems--supervisory ratings--is typical of systems used in other organizations, thereby providing an opportunity to compare the two. The peer rating system studied may differ from similar systems in other ways that may make it superior to peer evaluation per se. For example, it requires that panels contain members from more than one peer group. It also requires a consensus across criteria as opposed to merely averaging individual ratings or relying upon a majority rule of some kind.

Quite apart from these reasons, however, peer evaluations may be superior because of inherent problems in supervisory ratings. Supervisory ratings may be subject to range restriction, since supervisors are not given as much flexibility in their rating scales as are

peer raters. That is, a supervisor may feel constrained to give a certain rating value to a researcher doing an acceptable job at a given grade level; peer raters may have a larger range of ratings that pertain to these same circumstances. A supervisor may also restrict rating ranges by using cohorts of the researchers as a comparison standard. The same number of publications, for example, may be seen as evidence of exceptionally high performance for a lower grade scientist but only marginal performance for scientists at higher organizational levels. This possibility was tested by holding grade constant while relating the mean 1975 supervisor rating and the most recent total peer panel score to the linearly combined performance variable, "total." If the supervisors rated as suggested, the partialled relationship should increase for the supervisor rating. Conversely, if a grade bias inflated either peer or supervisor ratings, the respective partialled relationship should decrease. The partial coefficients for supervisor rating and peer rating with the performance variable were -.02(N.S) and .48 ($p < .001$), respectively. These findings suggest that the supervisors were not making within-grade discriminations in evaluating performance and that such ratings appear to be highly contaminated by the grade of the scientist. It appears that supervisory ratings could be based solely on the grade of the ratee as evidenced by the bias found in these evaluations. By comparison, only a minimal amount of grade-related bias was found among peer ratings.

While there are a number of similarities between the results of this research and those reviewed by Kane and Lawler (1978), there are interesting differences that deserve further comment.

Kane and Lawler noted that the internal consistency reliability of peer ratings for the studies they reviewed was "woefully low" and suggested that this condition was inherent in the rating process. The results of the present research demonstrate that improvements in reliability can be achieved.

Kane and Lawler also questioned the intergroup reliability of peer ratings as opposed to peer nomination. If a peer evaluation is to be fair, it must provide equitable judgments across different disciplines. Our results indicate that there is no statistical evidence to support the criticism that one or more peer groups received favored treatment based on peer evaluations. It would be interesting to determine whether this equity exists among supervisory ratings. Such a difference may be related to the organizational visibility of each rating system. For the organization studied, and probably for others, the supervisory rating system was not created to monitor or account for peer group differences; such considerations, of course, are central to the peer evaluation process. In short, supervisory ratings are not subject to as much scrutiny as peer ratings.

Kane and Lawler suggested that peer ratings can be expected to have lower validities than peer nominations. The present results suggest that peer ratings can also have large validities, actually larger than the average peer nomination reported by Kane and Lawler. Considering that these peer ratings were used administratively for promotion decisions and not just for research, this result is impressive.

On the issue of bias, Kane and Lawler suggested that peer ratings are subject to a high degree of rater bias. Nevertheless, when compared to supervisory ratings, which ironically have often been the sole criterion of validity, the peer ratings in this study look much less biased. In fact, the supervisor ratings in this research appear to be almost entirely attributable to bias generated by the individual's pay grade. Of concern to any organization would be whether or not an individual's performance was related to his or her performance ratings. For a peer evaluation system to be regarded as efficient, it would be necessary to show that persons who attain different organizational levels via this process also demonstrate different performance levels. The present study showed that

moderate to strong significant relationships exist between the grade level of the research scientists and several measures of their performance.

Finally, Kane and Lawler suggested that little can be said about user reaction to peer ratings. It is important to understand that the system described here was not developed outside the organization, nor was it constructed unilaterally by nonresearchers. A need for an evaluation process sensitive to the unique properties of research positions was recognized by scientists and personnel specialists in this organization. This joint concern led to their combined efforts to build an efficient evaluation process that would be acceptable to both groups. The success of this system (which has been in use for over 18 years) can be traced, in large part, directly to these combined contributions. These facts, along with the realization that the peer rating system was developed by the scientists themselves in this organization, speak strongly for its acceptance. With most organizations, the acceptance of such a process might be regarded as an end in itself. Being a Federal agency, this organization must also be concerned with a regulatory agency, the Civil Service Commission. The Civil Service Commission, which is responsible for overseeing the employment of Government personnel, has examined this system and concluded that it yields acceptable personnel decisions.

The differences between Kane and Lawler's findings and this research should naturally raise the question of "why." While this question demands greater attention than can be given here, and while any answer given must be labeled speculative, one characteristic of this peer rating system that is probably responsible for its favorable comparison is the degree of structure in the system. The procedures and methods used by the peer panels are formally defined and have been standardized to a large degree. This structure no doubt has an effect upon the raters, training them in the process of evaluation, while the formation of panels serves as motivation to follow the procedure carefully.

Judging by the lack of applications of peer evaluation systems in various organizational settings, acceptance of such a process must be difficult to gain. The data presented in this report support the argument that such an evaluation procedure measures consistently, measures accurately, and does both of these things better than a fairly standard supervisory rating system.

A final point can be made regarding the usefulness of peer evaluation contrasted with supervisory evaluations. It is important for an organization to be able to justify its personnel decisions. The results of this study show that the peer ratings are more closely related to scientific productivity than are supervisory ratings, and organizations weighing the costs and benefits of relying upon the peer evaluation system for personnel decisions should take this into account. Whenever individual performance on objective performance criteria is related to evaluation and high evaluation leads to organizationally administered rewards, the evaluation system is likely to encourage high productivity and performance.

RECOMMENDATIONS

Few opportunities exist to compare alternate evaluation systems; even fewer, to examine a peer rating system that affects personnel actions. This research has done both, and the results suggest that there are many positive features pertaining to peer ratings that recommend their inclusion in the performance evaluation process.

Peer ratings have common ground with supervisory ratings but show greater stability over time and are more highly related to objective measures of performance. Thus, organizations having to evaluate performance among individuals at levels similar to those in this study should consider the merits of peer ratings. Such a system is consistent, accurate, and fair, thus helping to form a firm basis for justifying personnel actions. Clearly, peer input promises useful information.

The perceived relationships between individual performance and organizational rewards can be affected considerably by the method of performance evaluation that the organization uses. Since peer ratings demonstrate a high relationship to objective measures of performance--higher than supervisory ratings--an organization could help stimulate greater effort among employees by adopting an evaluation process that is perceived as emphasizing more objective measures of productivity.

REFERENCES

Barrett, R. S., Taylor, E. K., Parker, J. W., & Martens, L. Scale information and supervisory ratings. Personnel Psychology, 1958, 11, 333-346.

Bernardin, H. J. Behavioral expectation scales versus summated scales: A fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 1976, 61, 364-370.

Borman, W. C., & Dunnette, M. D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Burnaska, R. F., & Hollmann, T. D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 1974, 59, 307-312.

DeCotiis, T. A. An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977, 19, 247-266.

Dunham, R. B., & Herman, J. B. Development of a female faces scale for measuring job satisfaction. Journal of Applied Psychology, 1975, 60, 629-631.

Dunnette, M. D. A note on the criterion. Journal of Applied Psychology, 1963, 47, 251-254.

Kane, J. S., & Lawler, E. E., III. Methods of peer assessment. Psychological Bulletin, 1978, 85, 555-586.

Korman, A. K. The prediction of managerial performance: A review. Personnel Psychology, 1968, 21, 295-322.

Kunin, T. The construction of a new type of attitude measure. Personnel Psychology, 1955, 8, 65-78.

Lawani, S. M. Citation analysis and the quality of scientific productivity. Biosciences, 1977, 27, 26-31.

Lawler, E. E., III. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 369-381.

Lewin, A. Y., & Zwany, A. Peer nominations: A model, literature critique, and a paradigm for research. Personnel Psychology, 1976, 29, 423-447.

Locke, E. A., Smith, P. C., Kendall, L. M., Hulin, C. L., & Miller, A. M. Convergent and discriminant validity for areas and methods of rating job satisfaction. Journal of Applied Psychology, 1964, 48, 313-319.

Mayfield, E. C. Value of peer nominations in predicting life insurance sales performance. Journal of Applied Psychology, 1972, 56, 319-323.

McClure, G. E., Tupes, E. C., & Dailey, J. T. Research on criteria of officer effectiveness. Research Bulletin, May 1951, 51-58.

McKean, H. A., Mandel, J., & Steel, M. N. A rating scale method for evaluating research positions. Personnel Administration, 1960, (July/August), 29-36.

Roadman, H. E. An industrial use of peer ratings. Journal of Applied Psychology, 1964, 48, 211-214.

Schwab, D. P., Hennemann, H. G., III, & Decotiis, T. A. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Science Citation Index. Institute for Scientific Information. 325 Chestnut Street, Philadelphia, PA 19196.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology, Chicago: Rand McNally, 1976.

Smith, P. C., & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Smith, P. C., Kendall, L. M., & Hulin, C. L. The measurement of satisfaction in work and retirement. Chicago: Rand McNally, 1969.

Waters, L. K., & Waters, C. S. Peer nominations as predictors of short-term sales performance. Journal of Applied Psychology, 1970, 54, 42-44.

Weitz, J. Selecting supervisors with peer ratings. Personnel Psychology, 1958, 11, 25-35.

Williams, S. B., & Leavitt, H. J. Group opinion as a predictor of military leadership. Journal of Consulting Psychology, 1947, 11, 283-291.

Winer, B. J. Statistical principals in experimental design (2nd. Ed.). New York: McGraw-Hill, 1971.

REFERENCE NOTES

1. Handbook for using the research grade-evaluation guide in ARS. Agricultural Research Service, Personnel Division, USDA, Hyattsville, MD 20782.

2. Agricultural Research Service Administrative Memorandum 434-5, Evaluation plan for research positions and incumbents. Agricultural Research Service, Personnel Division, Hyattsville, MD 20782.

3. Civil Service Commission, CSC research grade-evaluation guide (TS-52), Washington, DC, June 1964.

DISTRIBUTION LIST

Chief of Naval Operations (OP-102) (2), (OP-11), (OP-14), (OP-987H)
Chief of Naval Research (Code 450) (4), (Code 458) (2)
Chief of Information (OI-2252)
Director of Navy Laboratories
Chief of Naval Education and Training (00A), (N-5)
Commander, Naval Military Personnel Command (NMPC-013C)
Commander, David W. Taylor Naval Ship Research and Development Center
Commander, Naval Air Development Center
Commander, Naval Ocean Systems Center
Commander, Naval Surface Weapons Center
Commander, Naval Weapons Center
Commanding Officer, Naval Training Equipment Center (Technical Library)
Commanding Officer, Naval Coastal Systems Center
Commanding Officer, Naval Underwater Systems Center
Commanding Officer, Naval Research Laboratory
Officer in Charge, Annapolis Laboratory, David W. Taylor Naval Ship Research and Development Center
Officer in Charge, White Oak Laboratory, Naval Surface Weapons Center
Officer in Charge, New London Laboratory, Naval Underwater Systems Center
Director, Training Analysis and Evaluation Group (TAEG)
Director, Naval Civilian Personnel Command
President, Naval War College
Provost, Naval Postgraduate School
Personnel Research Division, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base
Occupational and Manpower Research Division, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base
Technical Library, Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base
Flying Training Division, Air Force Human Resources Laboratory, Williams Air Force Base
CNET Liaison Office, Air Force Human Resources Laboratory, Williams Air Force Base
Technical Training Division, Air Force Human Resources Laboratory, Lowry Air Force Base
Advanced Systems Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base
Program Manager, Life Sciences Directorate, Air Force Office of Scientific Research (AFSC)
Army Research Institute for the Behavioral and Social Sciences (Reference Service)
Army Research Institute for the Behavioral and Social Sciences Field Unit--USAREUR (Library)
Military Assistant for Training and Personnel Technology, Office of the Under Secretary of Defense for Research and Engineering
Office of Personnel Management, Washington
Commandant, Industrial College of the Armed Forces
Science and Technology Division, Library of Congress
Defense Documentation Center (12)